Text Analysis:
A Crucial Part of Enterprise Data Initiatives

by Joseph Turian, Ph.D.

8/2013

AlchemyAPI™
Transforming Text Into Knowledge

This paper was underwritten by AlchemyAPI.

Executive Summary

- Experts estimate that Fortune 500 companies are losing $12 billion per year in value because they are not exploiting unstructured data such as text.
- Savvy organizations are capitalizing on the opportunity that many of their competitors are missing. They are "closing the loop" with data, using text analysis software on unstructured data as part of their big data initiatives.
- Text analysis software unlocks the hidden value in unstructured data, helping enterprise decision-makers to improve brand equity, increase revenue, and reduce operational costs.
- Organizations that realize there is value in text analysis, and want to use text analysis software, are faced with the buy-or-build dilemma.
- There are many advantages to buying mature, robust text analysis software such as AlchemyAPI, including predictable low costs, as well as reduced time-to-market.

What Is Text Analysis?

Text analysis is the process of adding structure to text. Once structured, data can be distilled into business intelligence. Text analysis generally means automatic text analysis, as opposed to the expensive and time-consuming approach of doing text analysis by hand. Although the manual approach is useful on a small scale, it doesn't scale to enterprise-size data sets.

"We are drowning in information, but we are starved for knowledge."—John Naisbitt, *The New York Times* bestselling author

For example, consider a media analysis firm that wants to understand media response to the iPhone 5. Every 24 hours, several hundred blog posts are published that contain the keywords "iphone 5." As part of their work, this company needs to know:

- How many of these blog posts are relevant and actually expressing opinions about iPhone 5?
- How many of these blog posts merely mention "iPhone 5" in passing?
- How many relevant blog posts are negative, and how many are positive?
- What particular aspects and features of iPhone 5 capabilities are being praised or criticized? e.g., pricing and screen size?
- For all of the above, what is the trend for the past six months?

It would be too expensive for the media analysis company to have humans read hundreds of blog posts daily for the last six months. With text analysis, computers do the work. Employing complex linguistic, statistical, and neural network algorithms, AlchemyAPI is able to read and understand text, at massive rates of speed. By using AlchemyAPI, the media analysis firm can easily generate reports to answer the questions above.

Illustration 1 shows an example news article in raw input form, while Table 1 shows structured output after text analysis.


The iPhone Takes to the Big Screen

By Walter S. Mossberg

On Its Thinnest and Fastest Phone Ever, Apple Offers 4-Inch Display, LTE Network

The world's most popular smartphone becomes significantly faster, thinner, and lighter this week, while a gaining a larger, 4-inch screen

Illustration 1

Author	Text Attribute	Value	Sentiment
Walter S. Mossberg	Entity	iPhone	Positive
Walter S. Mossberg	Entity	Apple	Positive
Walter S. Mossberg	Entity	Google Maps	Positive
Walter S. Mossberg	Keyword	price	Negative
Walter S. Mossberg	Keyword	new maps	Positive
Walter S. Mossberg	Keyword	screen	Positive

Table 1

Before text analysis, the original raw input format was not formatted in a way that would make answering the above questions simple. After text analysis has been performed, we have database-friendly structured data, such as entity-specific sentiment scores.

This newly structured data is easy to use by existing reporting and business intelligence software. Insights from the final reports can now be used for decision-making by the PR firm and their client.

Applications of Text Analysis

> "Listening—for brand mentions, complaints, and concerns—is the first element of any credible social engagement program. Businesses that listen can uncover sales opportunities, measure satisfaction, gauge reactions to marketing campaigns and message themes, uncover root causes behind events, and detect and respond to reputation and competitive threats." (Source: Alta Plana)

Text analysis has a broad array of applications within the enterprise, including brand-reputation management, market research, competitive intelligence, and customer service and support. In this section, we touch on enterprise applications at a high-level.

Voice of the customer can be learned from analyzing customer communications. These include internal documents such as call center logs, emails to support sites, and responses on survey forms.

Voice of the customer can also be gleaned from online reviews. By looking at customer comments on external sites, you can uncover previously unknown problems in customer relations, understand trends, and determine issue hotspots. It can also overlap with market research, to detect emerging topics of interest to your customers.

By looking at customer comments on external sites, you can uncover previously unknown problems in customer relations, understand trends, and determine issue hotspots.

Market research can benefit from text analysis by spotting emerging trends and discovering new markets. It can also include monitoring events that are of strategic interest to your organization. SemanticWeb.com suggests asking the following questions:

- What new deals have been won by your competition?
- Is the industry stable?
- Are your executives respected and how are they perceived in the news (positive/negative)?
- How do market participants view your competitors?
- What are the market trends?
- Which companies interact together?
- Turnover in key employees?

Your organization can also uncover sales opportunities. For example, people complaining about some problem they have—which your product solves—are untapped customers.

Media monitoring is the identification and analysis of mentions, whether by customers or journalists. These could be mentions of your company, your competitors, your products, and your competitors' products. "What are people saying about X on Twitter, on Facebook, in blog posts, and in blog comments? What kind of language are they using?" Social media posts can give a complete view of customer sentiment, especially with fine-grained questions such as: "Do people see my car as fast? Do they think my product is cool?"

Product launches should be accompanied by media monitoring to understand consumer reactions. In particular, because it is crucial to quickly detect emerging issues and head them off before they become more costly.

Sentiment analysis is a key component of many text analysis applications, including voice of the customer, market research, and media monitoring. Sentiment analysis should have the following properties:

- Accurate
- Scalable
- Fine-grained

Recommendation systems analyze and connect text to related articles, videos, or ads. Famously, Google developed the AdSense business model of showing relevant ads by using text analysis on webpage content. Media properties such as *The Huffington Post* commonly use recommendation systems to display related stories, thus keeping users on-site and clicking on more content. YouTube also uses text analysis as part of its Related Videos feature.

Historical perspective can be gleaned by digging into historical data, and discovering trends. Trend analysis can be done for all text analysis applications, including voice of the customer, market research, media monitoring, and sentiment. "How is my brand perception evolving over time? When have there been spikes of interest, both positive and negative? Why?"

Trend analysis can also be used to evaluate fluctuations on short time-scales, such as the minutes and hours after a major PR event. One of the major benefits of using a scalable text analysis solution such as AlchemyAPI is that it is inexpensive to run a full historical analysis, and the full historical analysis can be created in seconds.

Building reliable components is hard; it involves many hidden costs and total cost of ownership can be unpredictably high.

How to Perform Text Analysis

We have illustrated the potential of text analysis for savvy enterprise organizations, and have demonstrated actual examples of text analysis use in enterprise. How does one do text analysis? In the following sections, we illustrate common pitfalls, as well as best practices, for text analysis.

Pitfall: Raw Word Counts

One common pitfall is trying to perform text analysis using raw word counts. Although this approach is appealing because of its simplicity, it can't capture any context information and leads to unreliable components. For example:

- An article mentions "apple." Is it talking about the technology company, or the food? This distinction is particularly important for applications that involve brand recognition. Vision Critical explains: "Many companies face a very specific version of the disambiguation challenge: they've got a company or brand name that is a common English (or foreign-language) word, making it difficult to separate references to Avon Cosmetics from references to the Avon Theatre, Avon Indiana, and the TV character Avon Barksdale."
- Sentiment based upon raw word counts is unreliable. For example, negation ("This movie was not good") cannot be detected by raw word counts.

Reliable text analysis requires sophisticated techniques that include context-aware linguistic and statistical algorithms.

Pitfall: Home-Grown Text Analysis

There are two problems with building text analysis components in-house:

1. **Risk**: Building reliable components is hard; it involves many hidden costs and total cost of ownership can be unpredictably high. Besides unexpected cost, there is also the risk of project failure.
2. **Time to results**: It can take many months, if not years, to build high-quality, reliable text analysis components.

Risk and time-to-results in building text analysis components can be broken down into the following challenges:

- **Technical expertise:** Natural language processing (NLP) expert roles are hard to fill. NLP experts have to be able to get training data, build models, scale the solution, and keep it up-to-date. This requires a combination of expertise in statistics, linguistics, and software engineering. They might not have knowledge of how to build multilingual systems, if needed. Even well-funded organizations have great difficulty hiring the NLP talent that they desire.
- **Build time**: NLP systems can take many months if not years to engineer. For example, text on Twitter is particularly challenging, given the use of casual language, misspellings, slang, and bad grammar. In addition to coding time, there is also time required to train the components. Many models require weeks or months of training time before they can achieve state-of-the-art accuracy.
- **Scalability**: The system must be scaled to handle high-velocity and high-volume text. Usage requirements can be unpredictable. Major events cause a spike in social media activity. It can be operationally complex to scale up resources on demand, and subsequently to scale down resources when analytical demand subsides.
- **Maintenance**: The system must be kept up-and-running. Ideally, the system should be retrained periodically; language is evolving and new words and terms enter our lexicon. A reliable system that works today might produce less reliable results in a year.

Each of the above difficulties adds risk to developing text analysis components in-house. The compounded risk can lead to projects that are over budget, over time, or simply unsuccessful.

For most organizations, developing and maintaining NLP components is not a core competency, nor should it be. If you intend to build and maintain your own text analysis components, be sure to make a thorough assessment of the effort required, and the total cost of ownership.

APIs simplify the process of integrating text analysis and confer access to a broad range of software with a low initial investment.

Best Practice: Use Existing Best-of-Breed Components

The best practice for implementing text analysis is to integrate best-of-breed components from an established industry leader. Compared to building, the overall advantages of buying existing components are that costs are more transparent and that time-to-results is much faster.

One particularly attractive way of using existing text analysis components is through Application Programming Interfaces (APIs), which are pay-by-the-sip. APIs simplify the process of integrating text analysis and confer access to a broad range of software with a low initial investment.

In particular, the advantages of buying best-of-breed solutions using an API, instead of buying a software package, include:

- **Ease-of-use:** APIs are easy to plug into immediately.
- **Clear pricing structures:** Cost is metered based upon usage, and all costs are known up-front. There are no initial costs or maintenance expenses. Using an API shifts capital expenditure into an operational expenditure, and avoids the cash flow necessary for a large payment upfront. By comparison, buying software involves hidden costs such as hardware, installation, and maintenance.

Since APIs are flexible, easy to use, and have low initial cost, they are the lowest risk option. APIs also allow enterprises to generate value immediately, without a major investment.

Timing and execution speed can be key. APIs allow companies to engage in short term projects they otherwise couldn't. For example, if a customer needs text analysis on German, you can tell them to come back in several months; or you can use an API, and help them today.

The AlchemyAPI Advantage

Scale Advantage

The AlchemyAPI training set is around 250 times the size of Wikipedia. AlchemyAPI crawls tens of billions of web pages, including hundreds of millions of tweets, and their last crawl used over a thousand computers to collect and analyze all that data. This scale is extraordinary, and lends versatility to the AlchemyAPI product. It is also beyond the operating capacity of many companies.

Only by constantly analyzing social media and the web can anyone hope to keep up with the changes that are occurring to language on an ongoing basis. Language is evolving, and new words, idioms, and product names are created daily. To keep models current, one must continue to repeatedly crawl the web. AlchemyAPI recrawls tens of billions of web pages monthly; so the scale advantage is not just about the volume of text data on the web, but the volume at which it grows.

Cross-Domain Advantage

Many text analysis APIs have decided to focus on a very narrow piece of the pie. For example, some text analysis APIs focus primarily on the intelligence community and anti-terrorism. Others focus solely on the voice of the customer. AlchemyAPI, by comparison, is exposed to text from a wide variety of different types and domains. Their customer base is broad, spanning over ten different industry verticals—everything from contextual advertising to business intelligence to financial services applications. Additionally, AlchemyAPI's system has a built-in self-improvement capability, which allows it to generalize beyond narrowly-defined text analysis systems.

Human Advantage

The creation of text analysis software requires extensive man hours, which is used at the onset to annotate labeled training data. This training data is used to help the software achieve a certain threshold of accuracy. Those human annotators often require a high degree of sophistication, such as a degree in linguistics and a significant amount of training. Hiring people to create training data is expensive, but it is necessary to get a text analysis system off the ground. AlchemyAPI has invested many person-years of labor for annotation and refining their product. This is an expensive barrier for companies to get over if they want to build their own text analysis software, and illustrates another advantage to using an existing system.

Outlook

Most enterprise data initiatives are focused on exploiting data that is already structured, but a lot of the potential in big data lies in the 80% that is unstructured. However, unstructured data is hard to use. One organization's missed opportunity is another organization's competitive advantage; by using text analysis software and unlocking the potential in unstructured data, smart organizations are capturing value that their competitors do not. They are leveraging unstructured data from both inside and outside their organization.

One organization's missed opportunity is another organization's competitive advantage; by using text analysis software and unlocking the potential in unstructured data, smart organizations are capturing value that their competitors do not.

The first step to unlocking the hidden value in unstructured data is to use text analysis to add structure. Text analysis exposes the information in unstructured data, rendering it accessible to more common enterprise data software such as databases and business intelligence tools. Text analysis has a broad array of applications within the enterprise, including brand reputation management, market research, competitive intelligence, and customer service and support.

The best practice for implementing text analysis is to buy best-of-breed components from an established commercial entity. In particular, AlchemyAPI text analysis software is flexible, easy to use, and has low cost. AlchemyAPI has a scale advantage, cross-domain advantage, and human advantage, and those three things give them a higher degree of accuracy than most of the competition. By translating raw text into meaningful information, AlchemyAPI enables decision-makers to leverage unique comprehensive insights for competitive advantage.

About Joseph Turian

Joseph Turian, Ph.D., heads MetaOptimize LLC, which consults on data science, NLP, and machine learning. He also runs the MetaOptimize Q&A site, where Machine Learning and Natural Language Processing experts share their knowledge. He specializes in large data sets.

Joseph Turian holds a Ph.D. in computer science (with a focus on Machine Learning and Natural Language Processing) from New York University since 2007. During his graduate studies, he developed a fast, large-scale machine learning method for parsing natural language. He received his AB from Harvard University in 2001.

As a scientist, Joseph Turian has over 14 refereed publications in top NLP + ML conferences. His team submitted the best parser in EVALITA 2009 Main+Pilot tasks. He is an advocate for open-notebook science, releasing his research code on his github, and for broader scientific collaboration through the internet.

About AlchemyAPI

The product of over 75 person years of engineering effort, AlchemyAPI is a text mining platform providing the most comprehensive set of semantic analysis capabilities in the natural language processing field. Used over 3 billion times every month, AlchemyAPI enables customers to perform large-scale social media monitoring, target advertisements more effectively, track influencers and sentiment within the media, automate content aggregation and recommendation, make more accurate stock trading decisions, enhance business and government intelligence systems, and create smarter applications and services. If you would like to learn more about our company and services, please call us at 1-877-253-0308 or email info@alchemyapi.com.